Generalized Multivariate Rank Type Test Statistics via Spatial U-Quantiles

Weihua Zhou[1]

University of North Carolina at Charlotte

and

Robert Serfling[2]

University of Texas at Dallas

Final revision for *Statistics and Probability Letters*

May 2007

[1]Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA. Email: wzhou2@email.uncc.edu.

[2]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75083-0688, USA. Email: serfling@utdallas.edu. Website: www.utdallas.edu/~serfling.

Abstract

The classical univariate sign and signed rank tests for location have been extended over the years to the multivariate setting, including recent robust rotation invariant "spatial" versions. Here we introduce a broad class of rotation invariant multivariate spatial generalized rank type test statistics. For a given inference problem not restricted to location, the test statistics are linked through Bahadur-Kiefer representations with spatial median estimators in appropriately matched U-quantile location models. Under null and contiguous alternative hypotheses, related quadratic form statistics have central and noncentral chi-square limit distributions. Robustness properties in terms of breakdown points and influence functions of the associated estimators are quite favorable. Illustrative applications cover location, multivariate dispersion, and multiple linear regression.

AMS 2000 Subject Classification: Primary 62G10 Secondary 62H99.

Key words and phrases: hypothesis tests; multivariate analysis; nonparametric; generalized ranks; spatial quantiles; multiple regression; multivariate dispersion.

1 Introduction

Two classical procedures for testing univariate location are the sign and signed rank statistics. These are formally linked with the sample median and the Hodges-Lehmann location estimator via Bahadur representations [1], [6], unifying the convergence theory for tests and estimators. For general background on various extensions of these classical tests to the multivariate location setting, see Hettmansperger and McKean (1998). Pertinent to the present paper are the robust and rotation invariant "spatial" multivariate sign and signed rank tests introduced by Möttönen and Oja (1995) and further studied in [11]. For these test statistics, the respective associated multivariate location estimators happen to be the "spatial" multivariate median introduced independently by Dudley and Koltchinskii (1992) and Chaudhuri (1996) and the spatial multivariate Hodges-Lehmann estimator (spatial median of pairwise averages) introduced by Chaudhuri (1992). The latter paper in fact treats a family of spatial multivariate Hodges-Lehmann location estimators defined as the spatial median of m-wise averages, for each choice of $m = 1, 2, \ldots$ Recently, Möttönen, Oja and Serfling (2004) investigated the series of multivariate spatial signed rank methods linked with this family of estimators and showed that as a group they offer attractive tradeoffs between robustness and efficiency, competing favorably with the classical Hotelling T^2 test.

Here we introduce a broad class of rotation invariant multivariate spatial generalized rank test statistics, encompassing not only the above-mentioned tests for multivariate location problems but also diverse other settings, as noted below. These statistics are given by *sample multivariate centered rank functions* which are defined as inverses of *sample multivariate spatial U-quantile functions*, which recently have been formulated and studied by Zhou and Serfling (2007), who establish a Bahadur-Kiefer representation for multivariate spatial U-quantiles in which the leading terms are the centered rank functions. The relevant background on spatial U-quantiles that we need here is provided in Section 2.

In Section 3, for convenient quadratic form statistics associated with these generalized rank type test statistics, we give limiting central and noncentral chi-square distributions under null and alternative hypotheses. Robustness properties are explored in Section 4, in terms of breakdown and influence function properties of the associated multivariate location estimators. It is seen that the favorable robustness of sample spatial quantiles carries over to spatial U-quantiles and their associated generalized rank test statistics.

As illustrative applications, we treat location, multiple regression, and multivariate dispersion problems in Section 5. For example, we define a test procedure associated with the extension in [17] of the Theil-Sen nonparametric simple linear regression slope estimator to the setting of multiple linear regression.

In general, starting with a given testing problem, which itself need not be a location problem, we obtain tests for the original problem as the centered rank functions of spatial median estimators in a related U-quantile location model.

2 Spatial U-quantiles and corresponding Bahadur-Kiefer representations

Following Chaudhuri (1996), the *spatial quantile function* corresponding to a cdf F on $\mathbb{R}^d$ is defined over $\boldsymbol{u}$ in the open unit ball $\mathbb{B}^{d-1}$ as the d-vector $\boldsymbol{\theta} = \boldsymbol{Q}_F(\boldsymbol{u})$ which minimizes $E\{\Phi(\boldsymbol{u}, \boldsymbol{X} - \boldsymbol{\theta}) - \Phi(\boldsymbol{u}, \boldsymbol{X})\}$, with $\Phi(\boldsymbol{u}, \boldsymbol{t}) = \|\boldsymbol{t}\| + \langle \boldsymbol{u}, \boldsymbol{t} \rangle$, $\|\cdot\|$ the usual Euclidean norm, and $\langle \cdot, \cdot \rangle$ the usual Euclidean inner product. Here $\boldsymbol{Q}_F(\boldsymbol{0})$ is the well-known *spatial median*. Equivalently, in terms of the *spatial sign function* (or *unit vector function*),

$$\boldsymbol{S}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \ \boldsymbol{x} \in \mathbb{R}^d,$$

the quantile $\boldsymbol{Q}_F(\boldsymbol{u})$ at $\boldsymbol{u}$ may be represented as the solution $\boldsymbol{x}$ of

$$E\{\boldsymbol{S}(\boldsymbol{x} - \boldsymbol{X})\} = \boldsymbol{u}. \tag{1}$$

Thus the quantile function $\boldsymbol{Q}_F(\cdot)$ has an *inverse*, given at each point $\boldsymbol{x} \in \mathbb{R}^d$ by the point $\boldsymbol{u}$ in $\mathbb{B}^{d-1}$ for which $\boldsymbol{x}$ has a quantile interpretation as $\boldsymbol{Q}_F(\boldsymbol{u})$, that is, by $\boldsymbol{Q}_F^{-1}(\boldsymbol{x}) = E\{\boldsymbol{S}(\boldsymbol{x} - \boldsymbol{X})\}$, the *expected direction* to $\boldsymbol{x}$ from a random point $\boldsymbol{X} \sim F$. The function $\boldsymbol{Q}_F^{-1}(\boldsymbol{x})$ is also known as the *spatial centered rank function* [9]. Thus the spatial quantile function and spatial centered rank function are simply inverses of each other.

We now introduce *spatial U-quantiles* following [17], where detailed treatment as well as background on univariate U-quantiles may be found. Consider i.i.d. observations $\{X_1, \ldots, X_n\}$ from a probability distribution P on any measurable space $(\mathbb{X}, \mathcal{A})$ and a *vector-valued* kernel $\boldsymbol{h}(x_1, \ldots, x_m)$ mapping $\mathbb{X}^m$ into $\mathbb{R}^d$. (The case described above corresponds to $\mathbb{X} = \mathbb{R}^d$, $m = 1$, and $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$.) Let $\boldsymbol{h}(X_1, \ldots, X_m)$ have cdf H on $\mathbb{R}^d$. By substituting a random kernel evaluation for $\boldsymbol{X}$ in (1), i.e., in terms of $\boldsymbol{Y} \sim H$ and the sign function $\boldsymbol{S}(\cdot)$ on $\mathbb{R}^d$, the *spatial U-quantile function* corresponding to H is defined as the solution $\boldsymbol{y} = \boldsymbol{Q}_H(\boldsymbol{u})$ of the equation

$$E\{\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Y})\} = \boldsymbol{u}, \tag{2}$$

and the corresponding inverse function is $\boldsymbol{Q}_H^{-1}(\boldsymbol{y}) = E\{\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Y})\}$, which thus is the *spatial centered rank function* corresponding to H. Let H satisfy

(i) H has a density bounded on bounded subsets of $\mathbb{R}^d$.

(ii) If $d \geq 2$, H is not concentrated on a line.

Then for any $\boldsymbol{u}$ the solution $\boldsymbol{Q}_H(\boldsymbol{u})$ to (2) always exists and is unique.

Associated with H we define a natural *empirical cdf* H_n by placing equal probability mass on the $n_{(m)} = n(n-1)\cdots(n-m+1)$ kernel evaluations $\boldsymbol{h}(X_{i_1}, \ldots, X_{i_m})$ taken over all m-tuples $(i_1, \ldots, i_m)$ of distinct indices chosen from $\{1, \ldots, n\}$. (For $\boldsymbol{h}$ symmetric under permutation of its arguments, it suffices to define H_n by placing equal probability mass simply on the $\binom{n}{m}$ kernel evaluations $\boldsymbol{h}(X_{i_1}, \ldots, X_{i_m})$ taken over all m-sets $\{i_1, \ldots, i_m\}$ of distinct indices chosen from $\{1, \ldots, n\}$. The kernels in the examples of Section 5 are of this type.) The corresponding sample spatial centered rank function is the vector-valued U-statistic

$$\boldsymbol{Q}_{H_n}^{-1}(\boldsymbol{y}) = n_{(m)}^{-1} \sum \boldsymbol{S}(\boldsymbol{y} - \boldsymbol{h}(X_{i_1}, \ldots, X_{i_m})),$$

and, accordingly, the sample analogue of $\boldsymbol{Q}_H(\boldsymbol{u})$ is given by the solution $\boldsymbol{y} = \boldsymbol{Q}_{H_n}(\boldsymbol{u})$ of the equation $\boldsymbol{Q}_{H_n}^{-1}(\boldsymbol{y}) = \boldsymbol{u}$. Although the sample spatial U-quantile function $\boldsymbol{Q}_{H_n}(\boldsymbol{u})$ is biased for $\boldsymbol{Q}_H(\boldsymbol{u})$ (just as univariate quantiles are biased), the corresponding sample spatial centered rank function is indeed unbiased (as in the univariate case):

$$E\{\boldsymbol{Q}_{H_n}^{-1}(\boldsymbol{y})\} = E\{\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{h}(X_{i_1}, \ldots, X_{i_m}))\} = E\{\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Y})\} = \boldsymbol{Q}_H^{-1}(\boldsymbol{y}). \tag{3}$$

In order now to state the relevant *Bahadur-Kiefer representation* for the sample spatial U-quantile function, we note that for the function $\|\boldsymbol{y}\|$ the gradient or first order derivative is given by the sign function $\boldsymbol{S}(\boldsymbol{y})$, and the $d \times d$ Hessian or second order derivative is given by

$$\boldsymbol{D}_2(\boldsymbol{y}) = \left\{ \frac{1}{\|\boldsymbol{y}\|} \left[\boldsymbol{I}_d - \frac{1}{\|\boldsymbol{y}\|^2} \boldsymbol{y}\boldsymbol{y}' \right] \right\}.$$

Under (i) the matrix

$$\boldsymbol{D}_1(\boldsymbol{y}) = E\left\{ \frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Y}) \right\} = E\{\boldsymbol{D}_2(\boldsymbol{y} - \boldsymbol{h}(X_1, \ldots, X_m))\} = E\{\boldsymbol{D}_2(\boldsymbol{y} - \boldsymbol{Y})\}$$

is positive definite. Let us also assume

> (iii) $\boldsymbol{Q}_H^{-1}(\boldsymbol{y})$ is continuously differentiable and $\boldsymbol{D}_1(\boldsymbol{y})$ is locally Lipschitz for $\boldsymbol{y}$ in an open set $\mathbb{V}$ in $\mathbb{R}^d$.

Then [17, Theorem 1.1] the sample spatial U-quantile function almost surely satisfies

$$\begin{aligned} &\boldsymbol{Q}_{H_n}(\boldsymbol{u}) - \boldsymbol{Q}_H(\boldsymbol{u}) \\ &= -[\boldsymbol{D}_1(\boldsymbol{Q}_H(\boldsymbol{u}))]^{-1} n_{(m)}^{-1} \sum [\boldsymbol{S}(Q_H(\boldsymbol{u}) - \boldsymbol{h}(X_{i_1}, \ldots, X_{i_m})) - \boldsymbol{u}] + \boldsymbol{R}_n(\boldsymbol{u}), \end{aligned} \tag{4}$$

with $\boldsymbol{R}_n(\boldsymbol{u})$ uniformly negligible (e.g., $o_p(n^{-1/2})$) over $\boldsymbol{u}$ in compact $K \subset \boldsymbol{Q}_H^{-1}(\mathbb{V}) \subset \mathbb{B}^{d-1}$.

3 Generalized Rank Type Test Statistics

We now formulate our generalized rank type test statistics and treat convergence properties. A central role is played by the Bahadur-Kiefer representation (4).

3.1 Formulation

As in Section 2, our setting is that of a sample of i.i.d. $\mathbb{X}$-valued observations $\{X_1, \ldots, X_n\}$, and a *vector-valued* kernel $\boldsymbol{h}(x_1, \ldots, x_m)$ mapping $\mathbb{X}^m$ into $\mathbb{R}^d$. Note that the *spatial median* of H is given by $\boldsymbol{Q}_H(\boldsymbol{u})$ with $\boldsymbol{u} = \boldsymbol{0}$, i.e., $\boldsymbol{Q}_H(\boldsymbol{0})$. We shall denote this parameter by $\boldsymbol{\theta}$. Our goal is to test a hypothesis of form $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Considering (4) with $\boldsymbol{u} = \boldsymbol{0}$, the *left hand side* thus represents the error in estimating the unknown location parameter $\boldsymbol{\theta}$ by $\boldsymbol{Q}_{H_n}(\boldsymbol{0})$, the spatial median of H_n. Hence a natural test of H_0

is given by comparing $\boldsymbol{Q}_{H_n}(\mathbf{0})$ with $\boldsymbol{\theta}_0$. Via the *right hand side* of (4), we equivalently may use (ignoring a constant factor and a minus sign) the test statistic

$$\boldsymbol{W}_n = n_{(m)}^{-1} \sum \boldsymbol{S}(\boldsymbol{\theta}_0 - \boldsymbol{h}(X_{i_1}, \ldots, X_{i_m})), \tag{5}$$

which we note is simply the relevant sample spatial centered rank function evaluated at $\boldsymbol{\theta}_0$, i.e., $\boldsymbol{W}_n = \boldsymbol{Q}_{H_n}^{-1}(\boldsymbol{\theta}_0)$. Indeed, this fact in itself provides a natural motivation for $\boldsymbol{W}_n$ as a test statistic, since under H_0 we have $\boldsymbol{Q}_H^{-1}(\boldsymbol{\theta}_0) = \mathbf{0}$ and via (3) we have $E\{\boldsymbol{W}_n\} = \boldsymbol{Q}_H^{-1}(\boldsymbol{\theta}_0)$, so that H_0 may be tested by measuring the closeness of the sample analogue $\boldsymbol{Q}_{H_n}^{-1}(\boldsymbol{\theta}_0)$ to $\mathbf{0}$. Thus, while the Bahadur-Kiefer representation is a powerful illuminating tool, we do not need all of its underlying assumptions in order to arrive at $\boldsymbol{W}_n$ as a natural test procedure. In particular, the regularity assumption (iii) can be somewhat relaxed. Another attractive feature of $\boldsymbol{W}_n$ is that it is simply a *vector-valued U-statistic* in structure.

In what follows, we study the statistic $\boldsymbol{W}_n$ in the context of a *U-quantile location model*, for which purpose we introduce a further assumption on H:

(iv) $H(\boldsymbol{y}) = H_0(\boldsymbol{y} - \boldsymbol{\theta})$, with H_0 *centrally symmetric* about $\mathbf{0}$,

i.e., for $\boldsymbol{Y}_0 \sim H_0$, $\boldsymbol{Y}_0 \stackrel{d}{=} -\boldsymbol{Y}_0$. Hence H_0 has spatial median $\mathbf{0}$ and H spatial median $\boldsymbol{\theta}$ [15]. Also, in order to avoid cumbersome notation and burdensome details of exposition, we now assume that the kernel $\boldsymbol{h}$ satisfies

(v) $\boldsymbol{h}$ is invariant under permutations of its arguments,

which is satisfied in typical examples. Under (v), the U-statistic $\boldsymbol{W}_n$ may be written as

$$\boldsymbol{W}_n = \binom{n}{m}^{-1} \sum \boldsymbol{S}(\boldsymbol{\theta}_0 - \boldsymbol{h}(X_{i_1}, \ldots, X_{i_m})), \tag{6}$$

which expression we shall adopt henceforth.

3.2 Null Hypothesis Results

Via standard results for U-statistics, $\boldsymbol{W}_n$ is found to be asymptotically d-variate normal and a suitable quadratic form asymptotically chi-square in distribution. In order to state the relevant parameters, we develop some details involving U-statistic projections. Unless otherwise indicated, expectations are in the U-quantile model corresponding to H with spatial median $\boldsymbol{\theta}$.

Define

$$\boldsymbol{K}(x, \boldsymbol{y}) = E_P\{\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{h}(x, X_1, \ldots, X_{m-1}))\}$$

and note that $E\{\boldsymbol{K}(X, \boldsymbol{y})\} = E\{\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{Y})\} = \boldsymbol{Q}_H^{-1}(\boldsymbol{y})$ ($= \mathbf{0}$ when $\boldsymbol{y} = \boldsymbol{\theta}$). In the $\boldsymbol{\theta}$ model, the projection (e.g., [13, Section 5.3.1]) of the U-statistic $\boldsymbol{W}_n$ is given by

$$\widehat{\boldsymbol{W}}_n = \sum_{i=1}^{n} E\{\boldsymbol{W}_n | X_i\} - (n-1) E\{\boldsymbol{S}(\boldsymbol{\theta}_0 - \boldsymbol{h}(X_{i_1}, \ldots, X_{i_m}))\}$$

and we obtain

$$\widehat{\boldsymbol{W}}_n - \boldsymbol{Q}_H^{-1}(\boldsymbol{\theta}_0) = \frac{m}{n}\sum_{i=1}^{n}[\boldsymbol{K}(X_i, \boldsymbol{\theta}_0) - \boldsymbol{Q}_H^{-1}(\boldsymbol{\theta}_0)]. \tag{7}$$

By [13, Section 5.3.4] we readily establish

$$\sqrt{n}\|\boldsymbol{W}_n - \widehat{\boldsymbol{W}}_n\| \xrightarrow{P} 0. \tag{8}$$

The asymptotic distribution theory of $\boldsymbol{W}_n$ is thus equivalent to that of $\widehat{\boldsymbol{W}}_n$. We will use the matrix

$$\boldsymbol{B}_{\boldsymbol{\theta}} = \mathrm{Cov}(\boldsymbol{K}(X, \boldsymbol{\theta}_0)) \quad \left(= E_{\boldsymbol{\theta}_0}\{\boldsymbol{K}(X, \boldsymbol{\theta}_0)\boldsymbol{K}(X, \boldsymbol{\theta}_0)'\} \text{ under } H_0\right),$$

which represents a *generalized rank covariance matrix* as for a special case in [10]. We thus have

Lemma 1

$$\sqrt{n}[\widehat{\boldsymbol{W}}_n - \boldsymbol{Q}_H^{-1}(\boldsymbol{\theta}_0)] \xrightarrow{d} N_d(\boldsymbol{0}; m^2\boldsymbol{B}_{\boldsymbol{\theta}}), \ \textit{as } n \to \infty. \tag{9}$$

In particular, under H_0, (7) becomes $\boldsymbol{W}_n = \frac{m}{n}\sum_{i=1}^{n}\boldsymbol{K}(X_i, \boldsymbol{\theta}_0)$ and (9) becomes

$$\sqrt{n}\,\widehat{\boldsymbol{W}}_n \xrightarrow{d} N_d(\boldsymbol{0}; m^2\boldsymbol{B}_{\boldsymbol{\theta}_0}).$$

By standard transformation results, this immediately yields

Theorem 2 *Under H_0, the limiting distribution of*

$$T_n = n\,m^{-2}\,\boldsymbol{W}_n'\boldsymbol{B}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{W}_n$$

is chi-square with d degrees of freedom.

A consistent estimator of $\boldsymbol{K}(x, \boldsymbol{\theta}_0)$ is given by

$$\widehat{\boldsymbol{K}}(x, \boldsymbol{\theta}_0) = \binom{n}{m-1}^{-1}\sum \boldsymbol{S}(\boldsymbol{\theta}_0 - \boldsymbol{h}(x, X_{i_1}, \ldots, X_{i_{m-1}})),$$

whence $\boldsymbol{B}_{\boldsymbol{\theta}_0}$ may be consistently estimated under H_0 by

$$\widehat{\boldsymbol{B}}_{\boldsymbol{\theta}_0} = n^{-1}\sum_{i=1}^{n}\widehat{\boldsymbol{K}}(X_i, \boldsymbol{\theta}_0)\widehat{\boldsymbol{K}}(X_i, \boldsymbol{\theta}_0)'.$$

Theorem 2 remains valid with substitution of this estimator for $\boldsymbol{B}_{\boldsymbol{\theta}_0}$ in T_n.

3.3 Alternative Hypotheses and Asymptotic Relative Efficiencies

Suppressing extensive details, we state convergence results for $\boldsymbol{W}_n$ under a sequence of contiguous alternative hypotheses $H^{(n)}$ represented by the parameter sequence

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \frac{\boldsymbol{\delta}}{\sqrt{n}}.$$

Under appropriate Pitman regularity conditions [7] the following result holds.

Theorem 3 *Under the sequence of alternatives $H^{(n)}$, the limiting distribution of T_n is noncentral chi-square with d degrees of freedom and noncentrality parameter*

$$\boldsymbol{\delta}'\, \boldsymbol{D}_1(\boldsymbol{\theta}_0)'\, \boldsymbol{B}_{\boldsymbol{\theta}_0}^{-1}\, \boldsymbol{D}_1(\boldsymbol{\theta}_0)\, \boldsymbol{\delta}.$$

Theorem 3 remains valid with $\widehat{\boldsymbol{B}}_{\boldsymbol{\theta}_0}$ substituted for $\boldsymbol{B}_{\boldsymbol{\theta}_0}$ in T_n. If another test statistic $\boldsymbol{V}_n$ satisfies the convergence result of Theorem 3 with noncentrality parameter

$$\boldsymbol{\delta}'\, \boldsymbol{C}\, \boldsymbol{\delta},$$

then the Pitman asymptotic relative efficiency of $\boldsymbol{W}_n$ with respect to $\boldsymbol{V}_n$ is the ratio of the respective noncentrality parameters,

$$\text{ARE} = \frac{\boldsymbol{\delta}'\, \boldsymbol{D}_1(\boldsymbol{\theta}_0)'\, \boldsymbol{B}_{\boldsymbol{\theta}_0}^{-1}\, \boldsymbol{D}_1(\boldsymbol{\theta}_0)\, \boldsymbol{\delta}}{\boldsymbol{\delta}'\, \boldsymbol{C}\, \boldsymbol{\delta}}.$$

While this ARE depends upon the direction of the alternatives from $\boldsymbol{\theta}_0$, the maximum and minimum eigenvalues of the matrix $\boldsymbol{D}_1(\boldsymbol{\theta}_0)'\, \boldsymbol{B}_{\boldsymbol{\theta}_0}^{-1}\, \boldsymbol{D}_1(\boldsymbol{\theta}_0)\, \boldsymbol{C}^{-1}$ provide useful bounded upper and lower bounds.

4 Generalized Location Estimators and Robustness Properties

Let us now consider estimation of the true unknown value of $\boldsymbol{\theta}$. Related to the test statistic $\boldsymbol{W}_n$ viewed as a function of $\boldsymbol{\theta}_0$ is the rotation equivariant location estimator $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{Q}_{H_n}(\boldsymbol{0})$, which is the sample analogue of the spatial median $\boldsymbol{Q}_H(\boldsymbol{0})$ of H and also is the solution of the equation

$$\boldsymbol{W}_n(\boldsymbol{\theta}_0) = 0.$$

Using the Bahadur-Kiefer representation (4) with $\boldsymbol{u} = \boldsymbol{0}$ and Lemma 1 with $\boldsymbol{\theta_0} = \boldsymbol{\theta}$, we obtain

Theorem 4

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\boldsymbol{0}, m^2 \boldsymbol{D}_1(\boldsymbol{\theta})^{-1} \boldsymbol{B}_{\boldsymbol{\theta}} \boldsymbol{D}_1(\boldsymbol{\theta})^{-1}).$$

The generalized variance $|\det \boldsymbol{D}_1(\boldsymbol{\theta})^{-1} \boldsymbol{B}_{\boldsymbol{\theta}} \boldsymbol{D}_1(\boldsymbol{\theta})^{-1}|$ plays a role in asymptotic relative efficiency comparisons with competing estimators, as described in [7] and [13].

One popular measure of the robustness is the *influence function*. It is straightfoward to show that the functional corresponding to $\widehat{\boldsymbol{\theta}}_n$ has IF given by the projection of the leading term in the Bahadur-Kiefer representation (4):

$$\mathrm{IF}(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}, x) = \boldsymbol{D}_1(\boldsymbol{\theta})^{-1}\boldsymbol{K}(x, \boldsymbol{\theta}),\ x \in \mathbb{X}.$$

Since $\|\boldsymbol{K}(x, \boldsymbol{\theta})\| \leq 1$, this IF is *bounded*. Another important measure of the robustness is the *breakdown point*. It is easily derived (see [12, p. 147] for discussion) that

$$\mathrm{BP}(\widehat{\boldsymbol{\theta}}_n) = 1 - (1/2)^{1/m}. \tag{10}$$

5 Applications

For different choices of kernel $\boldsymbol{h}(x_1, \ldots, x_m)$, the U-quantile approach yields competitive rank type test statistics with robust associated estimators. The scope of application of this approach, which defines the target parameter in the model of interest as a location parameter in a related U-quantile model, is quite broad. Here we examine three representative examples, covering not only location but also multiple regression and multivariate dispersion.

Example A *Generalized Hodges-Lehmann location inference.* Let $\mathbb{X} = \mathbb{R}^d$ and take

$$\boldsymbol{h}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \frac{\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_m}{m},$$

in which case $\boldsymbol{W}_n$ takes the form

$$\boldsymbol{W}_n = \binom{n}{m}^{-1} \sum \boldsymbol{S}(\boldsymbol{\theta}_0 - \boldsymbol{X}_{i_1} - \cdots - \boldsymbol{X}_{i_m}),$$

which is equivalent to the generalized spatial signed-rank test statistic of [10] (for $m = 1$ and 2 the spatial sign and signed rank test statistics, respectively, of [9]). The corresponding generalized Hodges-Lehmann location estimators

$$\widehat{\theta}_n^{(m)} = \text{spatial median}\left\{\frac{\boldsymbol{X}_{i_1} + \cdots + \boldsymbol{X}_{i_m}}{m}\right\}$$

for $m \geq 1$ are studied by Chaudhuri (1992). For F on $\mathbb{R}^d$ centrally symmetric about $\boldsymbol{\theta}$, the cdf H corresponding to $\boldsymbol{h}$ is also centrally symmetric about $\boldsymbol{\theta} = \boldsymbol{Q}_F(\boldsymbol{0}) = \boldsymbol{Q}_H(\boldsymbol{0})$. The same holds true, more generally, for the kernel

$$\boldsymbol{h}_{\boldsymbol{\alpha}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \sum_{i=1}^{m} \alpha_i\, \boldsymbol{x}_i,$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ satisfying $\sum_{i=1}^m \alpha_i = 1$ but otherwise unrestricted (although this kernel, however, is *not* symmetric in its arguments). Thus all of these test statistics and estimators for differing m and $\boldsymbol{\alpha}$ are competitors for inference about the same location parameter. The *univariate* case of $\boldsymbol{h}_{\boldsymbol{\alpha}}$ was introduced for $m = 2$ by Maritz, Wu and Staudte (1977) and treated for general m

by Choudhury and Serfling (1988). ■

Example B *Nonparametric inference on multiple regression slope coefficients.* Following [17], where detailed discussion may be found, consider the multiple linear regression model $Y = \alpha + \boldsymbol{\beta}'\boldsymbol{Z} + \varepsilon$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_p)'$, with i.i.d. observations $(\boldsymbol{Z}_i, Y_i, \varepsilon_i)$, $1 \le i \le n$, the errors $\{\varepsilon_i\}$ being independent of the random regressors $\{\boldsymbol{Z}_i\}$. Eliminate the parameter α by reducing the data to the $\binom{n}{2}$ differences

$$Y_i - Y_j = (\boldsymbol{Z}_i - \boldsymbol{Z}_j)'\boldsymbol{\beta} + \varepsilon_i - \varepsilon_j,\ 1 \le i < j \le n.$$

Let us denote the $\binom{n}{2}$ pairs (i, j) by $\mathbb{K}$. For each set K of p pairs $\{(i_1, j_1), \ldots, (i_p, j_p)\}$ from $\mathbb{K}$ with all indices $\{i_1, j_1, \ldots, i_p, j_p\}$ *distinct*, let $\boldsymbol{Y}_{(K)}$, $\boldsymbol{Z}_{(K)}$, and $\boldsymbol{\varepsilon}_{(K)}$, respectively, denote the p-vector of differences $Y_{i_m} - Y_{j_m}$, the $p \times p$ matrix of differences $\boldsymbol{Z}_{i_m} - \boldsymbol{Z}_{j_m}$, and the p-vector of differences $\varepsilon_{i_m} - \varepsilon_{j_m}$, for $m = 1, \ldots, p$. Thus

$$\boldsymbol{Y}_{(K)} = \boldsymbol{Z}_{(K)}{}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{(K)} \tag{11}$$

for each such K. We now define a relevant kernel as the least squares estimate of $\boldsymbol{\beta}$ based on equation (11). That is, for $K = \{(i_1, j_1), \ldots, (i_p, j_p)\}$,

$$\boldsymbol{h}((\boldsymbol{z}_{i_1}, y_{i_1}), (\boldsymbol{z}_{j_1}, y_{j_1}), \ldots, (\boldsymbol{z}_{i_p}, y_{i_p}), (\boldsymbol{z}_{j_p}, y_{j_p})) = (\boldsymbol{z}_{(K)}{}'\boldsymbol{z}_{(K)})^{-1}\boldsymbol{z}_{(K)}{}'\boldsymbol{y}_{(K)}. \tag{12}$$

By distinctness of the indices in K, each $\varepsilon_{i_m} - \varepsilon_{j_m}$ is a difference of independent and identically distributed observations and hence is symmetric about 0, and also the components of $\boldsymbol{\varepsilon}_{(K)}$ are independent. It easily follows that the vector $\boldsymbol{\varepsilon}_{(K)}$ in the model (11) has joint distribution *centrally symmetric* about the p-dimensional origin, i.e., $\boldsymbol{\varepsilon}_{(K)}$ and $-\boldsymbol{\varepsilon}_{(K)}$ are equal in distribution. This yields, that the random kernel evaluations have cdf in $\mathbb{R}^p$ *centrally symmetric* about $\boldsymbol{\beta}$, so that a natural nonparametric and robust estimator is thus given by

$$\hat{\boldsymbol{\beta}} = \text{spatial median}\{\boldsymbol{h}((\boldsymbol{Z}_{i_1}, Y_{i_1}), (\boldsymbol{Z}_{j_1}, Y_{j_1}), \ldots, (\boldsymbol{Z}_{i_p}, Y_{i_p}), (\boldsymbol{Z}_{j_p}, Y_{j_p}))\}, \tag{13}$$

which for the simple linear regression case $p = 1$ reduces to the classical estimator of Theil (1950). By (10), $\widehat{\boldsymbol{\beta}}$ is robust with BP equal to $1 - (1/2)^{1/(p+1)}$. Our Bahadur-Kiefer representation (4) and equation (6) provide an associated rotation invariant test statistic $\boldsymbol{W}_n$ whose corresponding quadratic form statistics have convenient limiting chi-square distributions. ■

Example C *Robust multivariate dispersion inference.* For F on $\mathbb{R}^d$, classical methods for inference about dispersion are based on the covariance matrix $\boldsymbol{\Sigma}$ and its sample analogue

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})',$$

which we note may be expressed equivalently as a U-statistic, $\binom{n}{2}^{-1}\sum_{1 \le i < j \le n}\boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{X}_j)$, based on the matrix-valued kernel $\boldsymbol{h}(\boldsymbol{x}_1, \boldsymbol{x}_1) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)(\boldsymbol{x}_1 - \boldsymbol{x}_2)'/2$, where $\boldsymbol{x}_1 = (x_{11}, x_{12})'$, etc. While unbiased for $\boldsymbol{\Sigma}$, the estimator $\boldsymbol{S}$ is not robust. Following [17], where detailed discussion may be found, we consider an alternative target parameter, $\boldsymbol{\Sigma}_{(2)}$ = the *spatial median* of the cdf H of

$\boldsymbol{h}(\boldsymbol{X}_1, \boldsymbol{X}_2)$, with robust sample analogue estimator $\boldsymbol{Q}_{H_n}(\mathbf{0})$. (The spatial median of the distribution of a random matrix $\boldsymbol{M}$ is defined as the usual spatial median of the distribution of vec $\boldsymbol{M}$.)

More generally, for integer $m \geq 2$, consider the matrix-valued kernel

$$\boldsymbol{h}^{(m)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \frac{1}{m-1} \sum_{i=1}^{m} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})' = \binom{m}{2}^{-1} \sum_{1 \leq i < j \leq m} \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

with kernel $\boldsymbol{h}$ as above. The U-statistic estimator based on $\boldsymbol{h}^{(m)}$ is unbiased for $\boldsymbol{\Sigma}$ but nonrobust. The alternative target parameter $\boldsymbol{\Sigma}_{(m)}$, however, defined as the *spatial median* of the cdf $H^{(m)}$ of $\boldsymbol{h}^{(m)}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$, has a robust sample analogue estimator $\boldsymbol{Q}_{H_n^{(m)}}(\mathbf{0})$, with BP equal to $1 - (1/2)^{1/m}$ independently of the dimension d.

For F a normal model in $\mathbb{R}^d$, $H_F^{(m)}$ is the Wishart$(\boldsymbol{\Sigma}, m-1)$ distribution with mean $\boldsymbol{\Sigma}$, and for this distribution we have that the spatial median $\boldsymbol{\Sigma}_{(m)} = c_m \boldsymbol{\Sigma}$ for some constant c_m. Thus, for each $m = 2, 3, \ldots$, a robust estimator for the usual covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\widetilde{\boldsymbol{\Sigma}}_{(m)} = c_m^{-1} \boldsymbol{Q}_{H_n^{(m)}}(\mathbf{0})$$

and related hypothesis testing may be carried out with the test statistic $\boldsymbol{W}^{(m)}$ corresponding to $\boldsymbol{Q}_{H_n^{(m)}}(\mathbf{0})$. For the univariate case this estimation approach has been investigated in detail in [14]. In robust *correlation* inference based on $\widetilde{\boldsymbol{\Sigma}}_{(m)}$ the constant c_m^{-1} conveniently becomes eliminated. ■

Acknowledgment

Support by NSF Grants DMS-0103698 and CCF-0430366 is gratefully acknowledged.

References

[1] Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics* **37** 577–580.

[2] Chaudhuri, P. (1992). Multivariate location estimation using extension of R-estimates through U-statistics type approach. *Annals of Statistics* **20** 897–916.

[3] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association* **91** 862–872.

[4] Choudhury, J. and Serfling, R. (1988). Generalized order statistics, Bahadur representations, and sequential nonparametric fixed-width confidence intervals. *Journal of Statistical Planning and Inference* **19** 269–282.

[5] Dudley, R. M. and Koltchinskii, V. I. (1992). The spatial quantiles. Preprint.

[6] Geertsema, J. C. (1970). Sequential confidence intervals based on rank tests. *Annals of Mathematical Statistics* **41** 1016–1026.

[7] Hettmansperger, T. P. and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods.* Arnold, London.

[8] Maritz, J. S., Wu, W., and Staudte, R. G. (1977). A location estimator based on a U-Statistic. *Annals of Statistics* **5** 779–786.

[9] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics* **5** 201–213.

[10] Möttönen, J., Oja, H., and Serfling, R. (2004). Multivariate generalized spatial signed-rank methods. *Journal of Statistical Research* **39** 25–48.

[11] Möttönen, J., Oja, H., and Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *Annals of Statistics* **25** 542–552.

[12] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley & Sons, New York.

[13] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

[14] Serfling, R. (2002). Efficient and robust fitting of lognormal distributions. *North American Actuarial Journal* **4** 95–109.

[15] Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference* **123** 259–278

[16] Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, III. *Proc. Kon. Ned. Akad. v. Wetensch. A* **53** 1397–1412.

[17] Zhou, W. and Serfling, R. (2007). Multivariate spatial U-quantiles: a Bahadur-Kiefer representation, a Theil-Sen estimator for multiple regression, and a robust dispersion estimator. *Journal of Statistical Planning and Inference*, to appear.